Peter Fankhauser, Norman Fiedler, Andreas Witt

Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik


Foto: privat
Peter Fankhauser


Foto: privat
Norman Fiedler


Foto: privat
Andreas Witt

Die Kernaufgabe des Instituts für Deutsche Sprache (IDS) ist die Erforschung und Dokumentation der deutschen Sprache. Dazu sammelt und archiviert das IDS einen umfangreichen Bestand an Forschungsprimärdaten in Form von Korpora der geschriebenen und gesprochenen Sprache sowie Sekundärdaten, wie zum Beispiel lexikographische Ressourcen. Dieser Beitrag gibt einen Überblick über den Datenbestand des IDS und die laufenden Forschungskooperationen im Bereich der Langzeitarchivierung. In diesem Kontext wird das im Aufbau befindliche Langzeitarchiv des IDS mit seiner Architektur, den zugrundeliegenden Prinzipien zur Daten- und Metadatenmodellierung sowie den daraus abgeleiteten Erfassungsprozessen vorgestellt. Der Beitrag schließt ab mit einem Ausblick auf die Herausforderungen und Perspektiven des Forschungsdatenmanagements aus Sicht der germanistischen Linguistik.

The core task of the Institute for the German Language (IDS) in Mannheim is the research into and documentation of the German language. IDS collects and preserves an extensive range of primary data in the form of corpora in both written and spoken language, as well as secondary data, such as lexicographical resources. This article provides an overview of the IDS collections and current cooperative research projects within the area of digital preservation. Further, it describes the architectural design of the IDS's digital archive, the fundamental principles underlying the data and metadata modeling, as well as the resulting data processing. In closing the article takes a look at the challenges and perspectives for managing research data from the point of view of German linguistics.

Einleitung

Warum Germanistik?

Von der Philosophie bis zu den Altertumswissenschaften, von der Anglistik bis hin zur Theologie bieten die Geisteswissenschaften bezüglich Themenvielfalt, Methoden und Quellen ein äußerst heterogenes Spektrum. Aber auch diese Disziplinen haben sich den Errungenschaften der modernen Informationsverarbeitung nicht verschlossen, sodass auch hier eine Vielzahl von digitalen Forschungsdaten anfällt. Um sich hinsichtlich der Klaviatur unterschiedlicher Formate und Datentypen nicht zu verirren, soll in diesem Beitrag die germanistische Sprachwissenschaft und pars pro toto das Institut für Deutsche Sprache (IDS) in Mannheim herausgegriffen werden, um Probleme und Lösungsansätze beim Forschungsdatenmanagement in den Geisteswissenschaften darzustellen.

Auch wenn das Forschungsmaterial des IDS nicht als prototypisch für die gesamten Sprachwissenschaften – geschweige denn die gesamten Geisteswissenschaften – angesehen werden kann, zeigen die Fragestellungen und Lösungswege zumindest exemplarisch Spezifika des Umgangs mit Forschungsdaten in den Geisteswissenschaften, die sich so in den Natur- und Lebenswissenschaften nicht finden.

Langzeitarchiv

Der Datenbestand des Instituts für Deutsche Sprache

Aufgrund seines Auftrages – der Erforschung und Dokumentation der deutschen Sprache in ihrem gegenwärtigen Gebrauch – hält das IDS relativ große Mengen an Forschungsprimärdaten vor, die für die empirisch arbeitende internationale Germanistik einen unschätzbar wichtigen Rohstoff für ihre Arbeit bilden. Ein großer Teil der Daten besteht aus sogenannten Sprachkorpora, d. h. aus digitalen Sammlungen von Daten des geschriebenen oder gesprochenen Deutsch unterschiedlicher Provenienz. Die Korpora liegen z. B. in Gestalt ganzer Texte geschriebener Sprache, als Audiodaten, multimodale Daten oder als versionierte Text-Korpora vor. Diese Forschungsdaten werden entweder selbst erhoben oder sie werden dem IDS von Textgebern, d. h. meist von Verlagen, für Forschungszwecke zur Verfügung gestellt. Neben den eigentlichen Daten umfasst der Datenbestand auch eine beachtliche Zahl von Metadaten, manuelle und automatische Annotationen sowie spezifische Zusammenstellungen von Sprachdaten. Diese Zusammenstellungen, die auch virtuelle Kollektionen genannt werden, liegen als Liste von Verweisen mittels persistenter Identifikatoren vor und erlauben den Linguistinnen und Linguisten spezifische Forschungsfragen, wie z. B. zum Sprachwandel, empirisch fundiert zu bearbeiten.

Der Bestand aller in das derzeit im Aufbau befindliche Langzeitarchiv des IDS eingespielten und dort bereitgehaltenen Forschungsdaten inklusive ihrer jeweiligen Metadaten beträgt derzeit etwa 1.000 Dateien von bis zu 32 GB in einem Gesamtumfang von 5 TB (Terabyte). Es ist jedoch davon auszugehen, dass der Umfang mit der zunehmenden Erschließung und Expansion des Datenbestandes in den nächsten Jahren schnell auf bis zu 10 TB an Daten des geschriebenen Deutschs ansteigen wird. Wie sich der künftige Bestand an multimodalen Daten, deren Zahl sich noch nicht einmal heute überblicken lässt, perspektivisch

entwickeln wird, kann zum jetzigen Zeitpunkt noch nicht seriös geschätzt werden.

Das IDS bemühte sich stets, seine Daten in standardisierten Formaten vorzuhalten, was zumindest bei der großen Mehrheit der Sprachdaten auch konsequent durchgehalten wurde. Aber auch die konsequente Verwendung von Standards führt zu einer recht großen Diversität von Datenformaten (vgl. Francopoulo 2006; Ide & Romary 2007; Stührenberg 2007, 2012; Riley & Becker 2009a; Riley & Becker 2009b; Romary 2011; Stührenberg et al. 2012). So werden die Metadaten z. B. im Format der TEI (s. www.tei-c.org) und als CMDI (s. Broeder et. al. 2011) gespeichert und die Korpora in dem älteren Standard XCES (vgl. Ide et al. 2000) oder in einem TEI P5-konformen Format (s. Lüngen & Sperberg-McQueen 2012).

standardisierte Formate

Die Forschungskomponente der E-Humanities

Im Rahmen der sogenannten E-Humanities oder Digital Humanities finden inzwischen zusehends Forschungsfragen Eingang in den Bereich des Forschungsdatenmanagements. So haben sich auch geisteswissenschaftliche Disziplinen und Forschungseinrichtungen zu Verbünden zusammengeschlossen, die gemeinsam zentrale Forschungsfragen adressieren.

Das europaweite Projekt *DARIAH* (www.dariah.eu) mit seinem deutschen Arm *DARIAH-DE* (de.dariah.eu) ist angetreten, die auf digitaler Basis arbeitenden Geisteswissenschaften mit Forschungsdaten und Werkzeugen zu ihrer Bearbeitung zu unterstützen. Darüber hinaus soll der Gedanke der E-Humanities in Forschung und Lehre hineingetragen werden, um die fachwissenschaftliche Nachhaltigkeit dieses Ansatzes mittels Ausbildung des wissenschaftlichen Nachwuchses zu gewährleisten. Ein Ziel hierbei ist es, die digitalen Geisteswissenschaften zu einer in den Naturwissenschaften üblichen Zusammenarbeit – basierend auf den vier Säulen Lehre, Forschung, Daten und Infrastruktur – zu bewegen.

Das Projekt *TextGrid* (www.textgrid.de) ist ein Forschungsverbund, der Geistes- und Kulturwissenschaften Zugang zu Informationen im Austausch und in gemeinschaftlicher Forschung mithilfe moderner digitaler Methoden der Datenerzeugung, -speicherung, -analyse und -edition ermöglichen möchte. Das seit 2006 vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Vorhaben widmet sich dem Problem, die Werkzeuge und Dienste für den Umgang mit Forschungsdaten in einer generischen und offenen Virtuellen Forschungsumgebung (VFU) zusammenzufassen. Darüber hinaus sollen Lösungen für einen organisatorisch, finanziell und wissenschaftlich nachhaltigen Betrieb solcher Plattformen gefunden werden.

Das Projekt *CLARIN-D* (de.clarin.eu) hingegen – es handelt sich auch hierbei um den deutschen Zweig eines gemeinsamen europäischen Vorhabens *CLARIN* (www.clarin.eu) – hat zahlreiche Kompetenzzentren auf nationaler Ebene aufgebaut und in einer webbasierten Infrastruktur zusammengefasst, um für die Geistes- und Sozialwissenschaften gemeinsame und standardisierte Daten, Datenformate und Werkzeuge in einer integrierten und interoperablen Plattform zu versammeln. Hierbei werden nicht nur die Anforderungen spezialisierter Teildisziplinen aufgegriffen, sondern auch Unterstützung zu übergreifenden Themen – zu nennen wären hier beispielsweise juristische Aspekte des Umgangs mit Sprachdaten – bereitgestellt.

Im Rahmen des Netzwerks *nestor* (www.langzeitarchivierung.de) tauschen sich schließlich zahlreiche Akteure aus Bibliotheken, Archiven, Museen und Forschungseinrichtungen zu Themen der Langzeitarchivierung (LZA) von Forschungsdaten aus. Diese Zusammenarbeit setzt auch auf Synergien bei Informationen, Standards und Aufgaben und zielt u. a. auch auf die Qualifikation der Geisteswissenschaftlerinnen und Geisteswissenschaftler bei der langfristigen Aufbewahrung ihrer Forschungsergebnisse ab. In *nestor* veröffentlichen aktive Mitglieder auch kontinuierlich grundlegende und umfassende Darstellungen zur Langzeitarchivierung von Forschungsdaten (vgl. nestor 2010, nestor 2012, Keitel & Schoger 2013).

Zusammenschluss zu Verbünden

Mit der Bündelung von (technischen) Serviceaufgaben in der Verwaltung, Pflege und Speicherung von Forschungsdaten, wie sie von EDV-Abteilungen und Bibliotheken durchgeführt werden, und mit einer ausgeprägten Forschungskomponente für Erstellung und Nutzung dieses empirischen Materials, unter anderem in Gestalt der beschriebenen Projekte, ist das IDS ein Vorreiter bei Aufbau und Betrieb von Forschungsinfrastrukturen in den Geisteswissenschaften und wurde auch im Sinne der Empfehlungen des Wissenschaftsrates (WissRat 2011) mehrfach erwähnt. Seit der Veröffentlichung dieser Empfehlungen hat das IDS konsequent den Ausbau des Programmbereichs »Forschungsinfrastrukturen« vorangetrieben, u. a. durch die organisatorische Integration der Forschungsbibliothek des IDS und der Arbeitsstelle »Zentrale Datenverarbeitung«. Im Folgenden sollen zentrale Aspekte der Arbeit mit linguistischen Forschungsdaten, die im Programmbereich »Forschungsinfrastrukturen« am Institut für Deutsche Sprache in Mannheim durchgeführt werden, präsentiert und auch verwandte Betätigungsfelder skizziert werden.

Bündelung von Serviceaufgaben

Das Forschungsdatenarchiv

Das IDS-Repositorium im Kontext

nachhaltige Archivierung von Sprachressourcen

Die Kernaufgabe des IDS-Repositoriums ist die nachhaltige Archivierung von Sprachressourcen. Abbildung 1 zeigt schematisch die Rolle des Repositoriums im Kontext der Ressourcen innerhalb und außerhalb des IDS. Die internen Ressourcen umfassen Korpora der geschriebenen Sprache (Deutsches Referenzkorpus, DeReKo), der gesprochenen Sprache (Archiv für Gesprochenes Deutsch, AGD) sowie lexikographische Ressourcen. Für die Aufbereitung und Nutzung dieser Ressourcen existieren jeweils eigene Aufbereitungsprozesse und Systeme, die ständig weiterentwickelt werden: Für geschriebene Sprache sind das derzeit das Korpus-Suche-, -Management- und -Analyse-System COSMAS II sowie die Aufbereitungsprozesse im Bereich DeReKo, für gesprochene Sprache die Datenbank Gesprochenes Deutsch (DGD) und für lexikographische Ressourcen das Online-Wortschatz-Informationssystem Deutsch (OWID). Das IDS-Repositorium dient also als zentrales, systemunabhängiges Langzeitarchiv für diese Systeme sowie für externe Ressourcen.

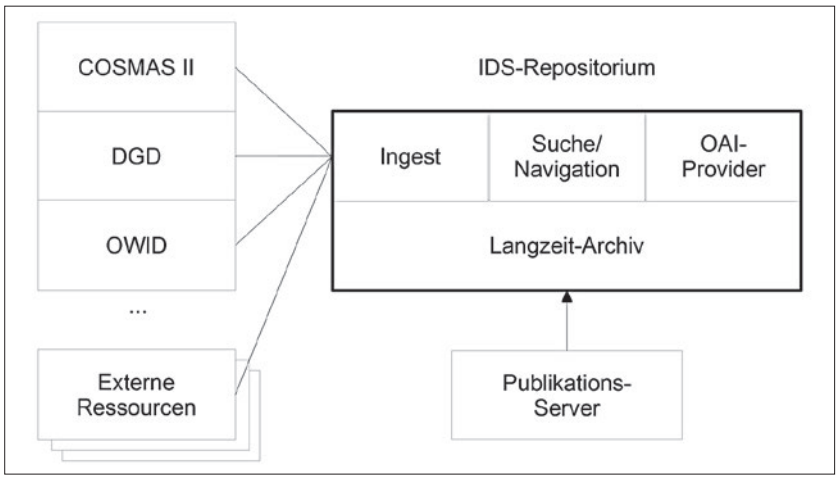

Abb. 1: IDS-Repositorium im Kontext

Funktionale Architektur

Abbildung 2 zeigt die funktionale Architektur des IDS-Repositoriums anhand des OAIS-Referenzmodells (OAIS 2012, Abschnitt 4.1). Die wesentlichen Komponenten wurden auf Basis von Fedora-Commons (Lagoze et al. 2005) realisiert.

»Archive Information Packages« (*AIPs*) verknüpfen Primärdaten mit deskriptiven und technischen Metadaten. Sie werden als digitale Objekte im Fedora-XML-Format (FOXML) serialisiert und gespeichert. Eine Ressource besteht typischerweise aus mehreren miteinander verknüpften digitalen Objekten. Jedes digitale Objekt besteht wiederum aus mehreren sogenannten Datenströmen für Daten und Metadaten in unterschiedlichen Formaten. Für alle Ressourcen wird eine Metadatenbeschreibung in CMDI (Broeder et al. 2011) sowie eine daraus abgeleitete Dublin-Core-Beschreibung abgelegt.

Die Einspeisung (*Ingest*) von »Submission Information Packages« (*SIPs*) wird von einem einfachen Webinterface sowie einem REST-API und Skripts für den Ingest von mehreren digitalen Objekten (Batch-Ingest) unterstützt. Das IDS-Repositorium verwendet typischerweise Batch-Ingest auf Basis von *SIPs*, die von einem für die jeweilige Ressource angepassten Verarbeitungsprozess (*FOXML-Generator*) erzeugt werden. Für die Suche unterstützt Fedora-Commons

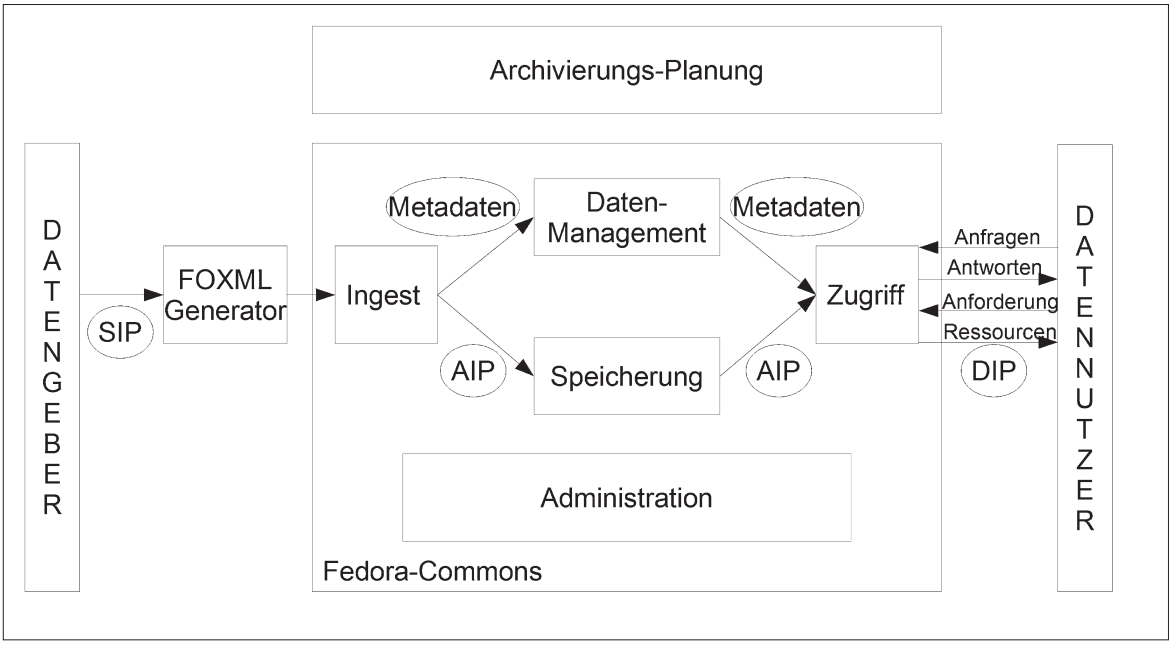

Abb. 2: Funktionale Architektur des IDS-Repositoriums

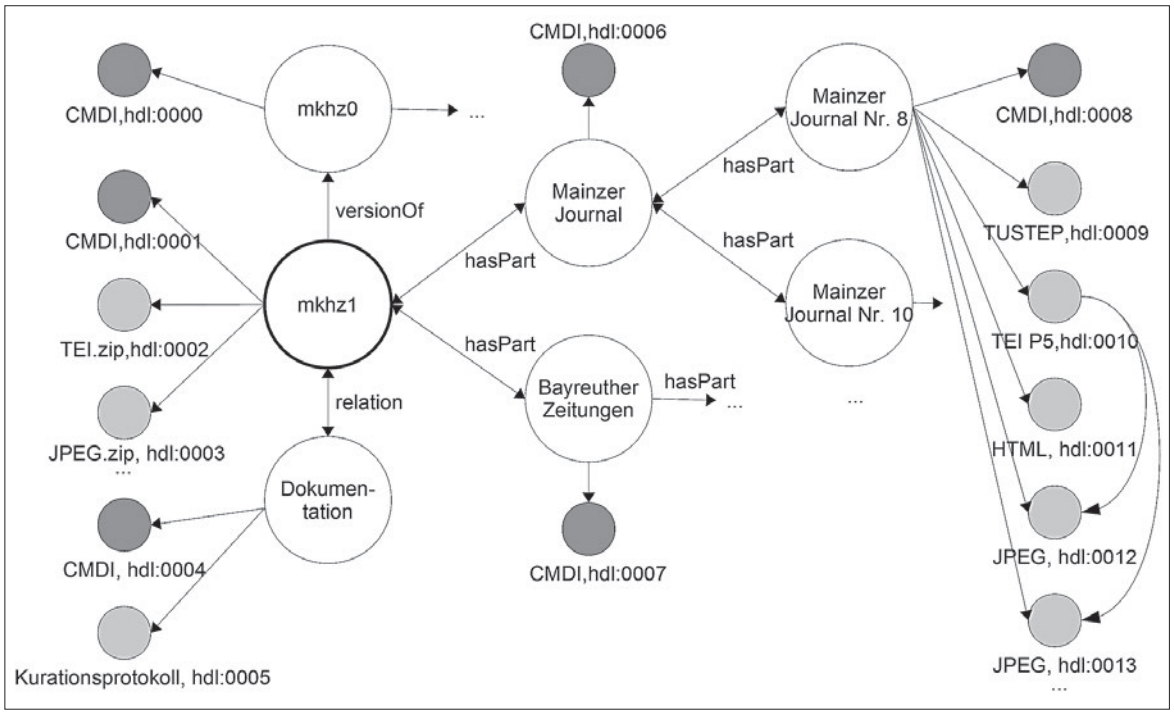

Abb. 3: Objektmodell

ein REST-API, das von einer einfachen formularbasierten Webschnittstelle angesprochen wird. Die Metadaten werden einerseits maschinenlesbar über eine OAI-PMH-konforme Schnittstelle (OAI-PMH) und anderseits in einer für Webbrowser geeigneten Darstellung zur Verfügung gestellt. Nutzerinnen und Nutzer haben direkten Zugriff auf die archivierten digitalen Objekte, wenn die geeignete Zugangsautorisierung besteht; einige Ressourcen und alle Metadaten werden ohne Zugangsbeschränkung zur Verfügung gestellt. Diese »Dissemination Information Packages« (*DIPs*) bestehen entweder aus einzelnen Datenströmen oder einem gepackten Archivformat für die gesamte Ressource.

Modellierungsprinzipien

Sprachressourcen sind sehr heterogen in Bezug auf verwendete Datenmodelle für Metadaten und Primärdaten, Medienformate, Speicherungssysteme sowie Umfang und Nutzungsform. Diese Heterogenität hat teilweise historische Gründe, ist aber auch auf die spezifischen Anforderungen und Bedingungen zurückzuführen, unter denen die Ressourcen zusammengestellt wurden. Ziel der Langzeitarchivierung ist es einerseits, die Ressourcen in ihrem aktuellen Nutzungskontext langfristig verfügbar zu machen, und andererseits, die Ressource auch in neuen Kontexten für eine Nachnutzung aufzubereiten. Dabei gelten folgende Prinzipien:

1. Erhaltung der Originalformate: Die Originaldaten werden in jedem Fall in ihrem Originalformat mit möglichst geringen Anpassungen (z. B. Kodierung in UTF-8) archiviert. Damit soll zumindest ihre aktuelle Nutzungsform gesichert sein.

Metadaten

2. Ergänzung der Originalformate um nachhaltige Formate: Falls ein Originalformat nicht einem der für die nachhaltige Archivierung und Nutzung empfohlenen Formate entspricht, werden die Originaldaten zusätzlich in geeignete Formate konvertiert und archiviert.
3. Erhaltung und Formalisierung der Metadaten: Alle vorhandenen Metadaten werden verlustfrei in ein formales CMDI-Profil übersetzt. Dieses Vorgehen beinhaltet die formale Abbildung der einzelnen Metadatenfelder in geeignete ISOcat-Kategorien.
4. Ergänzung von fehlenden Metadaten: Insbesondere für die Gesamtressource werden fehlende Kernmetadaten der Dublin-Core-Initiative (Ersteller, Titel, Beschreibung, Typ, Sprache) aus den Daten extrahiert oder manuell ergänzt.
5. Persistente Identifikation: Alle Ressourcen und ihre Datenströme erhalten einen sogenannten Persistenten Identifikator (*PID*) (Sun 2001), sodass sie auch bei einem Umzug des Repositoriums zu einer anderen Adresse unveränderlich referenzierbar bleiben.

Heterogenität der Sprachressourcen

Objektmodell

Digitale Sprachressourcen bestehen typischerweise aus mehreren Teilen. Abbildung 3 zeigt am Beispiel des Mannheimer Korpus Historischer Zeitungen und Zeitschriften (mkhz) das dem Repositorium zugrunde liegende Objektmodell.

Das Korpus (*mkhz1*) besteht aus mehreren Zeitungen (*Mainzer Journal, Bayreuther Zeitungen* etc.), die ihrerseits wieder aus mehreren Ausgaben bestehen. Diese Beziehungen werden als *hasPart* repräsentiert. Für jedes dieser Objekte werden Metadaten im CMDI-Standard angelegt, die Dublin-Core-Metadaten repräsentieren und dem Objekt einen persistenten Identifikator zuordnen – in Abbildung 3 exemplarisch dargestellt mit *hdl:nummer*. Die eigentlichen Daten werden als sogenannte Datenströme repräsentiert und ebenfalls mit einem persistenten Identifikator ausgestattet. So enthält das Objekt »*Mainzer Journal Nr. 8*« die einzelnen Druckseiten in hochaufgelöstem JPEG-Format, die ursprüngliche Transkription im TUSTEP-Format, die daraus generierte Transkription im TEI P5-Format und eine zum »Schmökern« geeignete Version in HTML. Die Transkriptionen verweisen ihrerseits wieder auf die ihnen zugrundeliegenden Druckseiten unter Verwendung derer persistenten Identifikatoren. Damit wird eine enge Verknüpfung zwischen den Primärdaten (Druckseiten) und den daraus abgeleiteten Sekundärdaten (Transkription) erreicht. Die einzelnen Formate (TUSTEP, TEI, JPEG) werden zusätzlich als komprimiertes Archiv (*.zip*) abgelegt und als Datenströme der Gesamtressource verfügbar gemacht. Das Objekt *Dokumentation* dient zur Archivierung der Aufbereitungsprozesse, und die Relation *versionOf* setzt das aktuelle Korpus in Bezug mit einer früheren Aufbereitung (*mkhzo*).

persistente Identifikatoren

Aufbereitungsschritte

Die anderen Korpora und Ressourcen des IDS verwenden eine ähnliche Modellierung, unterscheiden sich aber in den für das jeweilige Korpus adäquaten Metadaten, der gewählten Granularität und den zugrundeliegenden Formaten. Die Prinzipien der Repräsentation von Sprachressourcen für die Langzeitarchivierung – geeignete Granularität zur persistenten Identifizierbarkeit, Metadatenvollständigkeit zur Auffindbarkeit, Formatvollständigkeit zur nachhaltigen Nutzung sowie Dokumentation der Aufbereitung und aktuellen Nutzung zur Nachvollziehbarkeit – gelten jedoch für alle archivierten Ressourcen.

Metadatenmodellierung

Anders als zum Beispiel bibliographische Daten weisen Sprachressourcen eine hohe Heterogenität in ihren Metadaten auf. Korpora der geschriebenen Sprache erfordern andere Metadaten als Korpora der gesprochenen Sprache oder lexikographische Ressourcen. Um dieser Heterogenität Rechnung zu tragen, verwendet das IDS-Repositorium die in CLARIN entwickelte Metadaten-Infrastruktur CMDI (Broeder et al. 2011). CMDI ermöglicht es, beliebige Metadatenschemata auf Basis von wiederverwendbaren Komponenten zu spezifizieren. Die Semantik der Metadaten-Felder wird dabei über eine obligatorische Zuordnung zu einer ISOcat-Kategorie eindeutig spezifiziert. Durch die Wiederverwendung von Komponenten und die explizite Spezifikation der Semantik können so die spezifischen Metadaten-Anforderungen einer Sprachressource berücksichtigt, aber auch die Interoperabilität zwischen verschiedenen Metadaten-Schemata aufrechterhalten werden.

Metadaten-Infrastruktur CMDI

Wie bereits im Abschnitt *Modellierungsprinzipien* ausgeführt, wird für jede Ressource ein Minimum von Dublin-Core-Metadaten erhoben. Konkrete Beispiele der Metadaten sind im IDS-Repositorium einzusehen.

Ingestprozesse

Die Aufbereitung von Ressourcen erfordert häufig ressourcenspezifische Prozesse, die sich jedoch möglichst aus wiederverwendbaren und konfigurierbaren Komponenten zusammensetzen. Abbildung 4 zeigt eine verallgemeinerte Darstellung der wesentlichen Aufbereitungsschritte, die sich stark an den Empfehlungen im Leitfaden für Forschungsdaten-Management (Ludwig & Enke 2013, Kapitel 3) orientieren:

Im ersten Schritt (*Alinierung*) werden Metadaten und Daten in einen expliziten Bezug miteinander gesetzt. Metadaten werden häufig in Form von Tabellen, deren Zellen durch Kommata voneinander getrennt sind, oder durch ad hoc-XML-Strukturen mit Referenzen auf die eigentlichen Daten angegeben. Dieser Schritt stellt sicher, dass für jede Ressource Metadaten eindeutig vorhanden sind. Er erfordert häufig eine Normalisierung von Referenzen und Dateinamen.

Im zweiten Schritt (*Validierung/Kuration*) werden Datenformate validiert, und – falls erforderlich – nicht nachhaltig nutzbare Datenformate in eines der zur Langzeitarchivierung empfohlenen Datenformate konvertiert. Typischerweise werden dabei sowohl XML-basierte Formate (z. B. auf Basis von TEI P5) als auch PDF/A-Versionen erstellt. Nicht valide Dateien bzw. Duplikate werden in Abstimmung mit den Datengebern behandelt. Dieser Schritt zielt auf eine nachhaltige Nutzbarkeit der Daten. Abhängig vom Datenformat kann dieser Schritt recht aufwändig sein. Eine detaillierte Darstellung des Aufbereitungsprozesses für das Mannheimer Korpus Historischer Zeitungen und Zeitschriften ist in Fankhauser et al. 2013 verfügbar.

Im dritten Schritt (*Metadaten-Extraktion*) werden fehlende Metadaten aus der XML-Repräsentation der Daten extrahiert bzw. manuell hinzugefügt. Dieser Schritt nutzt die XML-Aufbereitung des vorherigen Schritts, um mithilfe von standardisierten XML-Werkzeugen weitere Metadaten zu erzeugen.

Im vierten Schritt (*CMDI-Generierung*) werden die Metadaten in ein geeignetes Profil der CMDI-Metadaten-Infrastruktur transformiert. Falls kein geeignetes Profil verfügbar ist, wird ein bestehendes Profil um entsprechende Metadatenfelder erweitert. Mit der Zuordnung von Metadatenfeldern zu formalen ISOcat-Kategorien zielt dieser Schritt auf eine formale Interpretierbarkeit der Metadaten.

Die Aufbereitungsschritte zielen jeweils auf einen Aspekt der Daten- bzw. Metadatenqualität, bauen modular aufeinander auf und werden umfassend dokumentiert. Das Ergebnis der Aufbereitung ist ein »Submission Information Package« (*SIP*), das aus mehreren digitalen Objekten (*Daten*) zusammen mit ihren *CMDI-Metadaten* besteht.

Im letzten Schritt werden diese *SIP*s für den Ingest in Fedora-Commons aufbereitet. Dieser Schritt ist als eine einfach konfigurierbare Verarbeitungskette realisiert und in Abbildung 5 schematisch dargestellt.

Zunächst werden CMDI-Metadaten mit fehlender Information ergänzt und gegen ihr Profil, das als XML-Schema repräsentiert ist, validiert. Daraufhin werden für jedes digitale Objekt, bestehend aus CMDI-Metadaten und Daten, FOXML-Datenströme generiert. Diese Generierung wird von FOXML-Schablonen gesteuert, die weitere technische Metadaten für Fedora-Commons bereitstellen, wie zum Beispiel Mime-Typ, lokale Identifikatoren und Kontrollgruppe. Zudem werden die von Fedora-Commons vorgegebenen Metadatenströme für Dublin Core und die OAI-PMH-Schnittstelle aus den CMDI-Metadaten mithilfe von XSLT-Stylesheets generiert (OAI-PMH). Diese Datenströme werden ebenfalls validiert. Für alle Datenströme wird ein persistenter Identifikator (*PID*) registriert sowie alle lokalen Identifikatoren im SIP mit dem zugehörigen PID ersetzt. Die Abbildung zwischen PIDs und lokalen Identifikatoren wird ebenfalls im Repositorium gespeichert. Die so generierten digitalen FOXML-Objekte werden in einem Batch-Prozess in Fedora-Commons eingespeist.

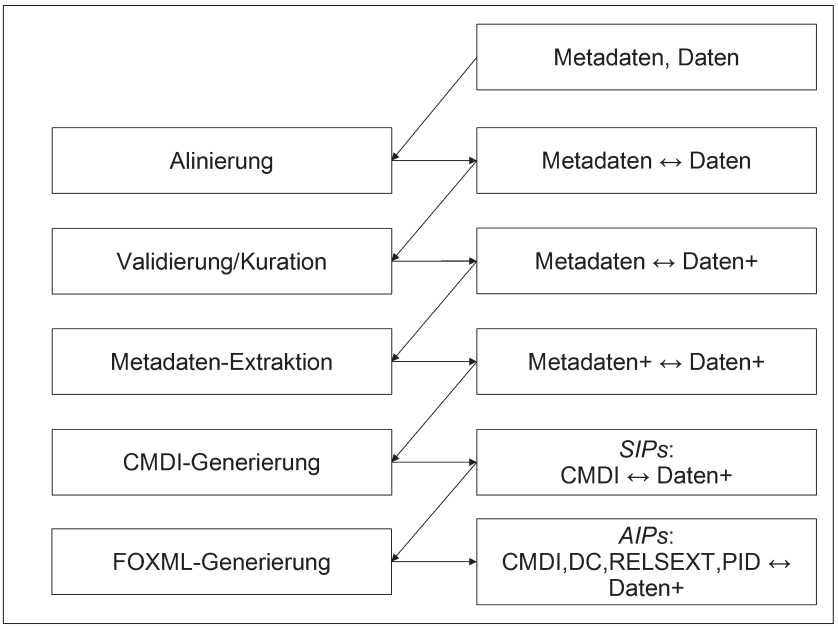

Abb. 4: Allgemeiner Aufbereitungsprozess von Sprachressourcen

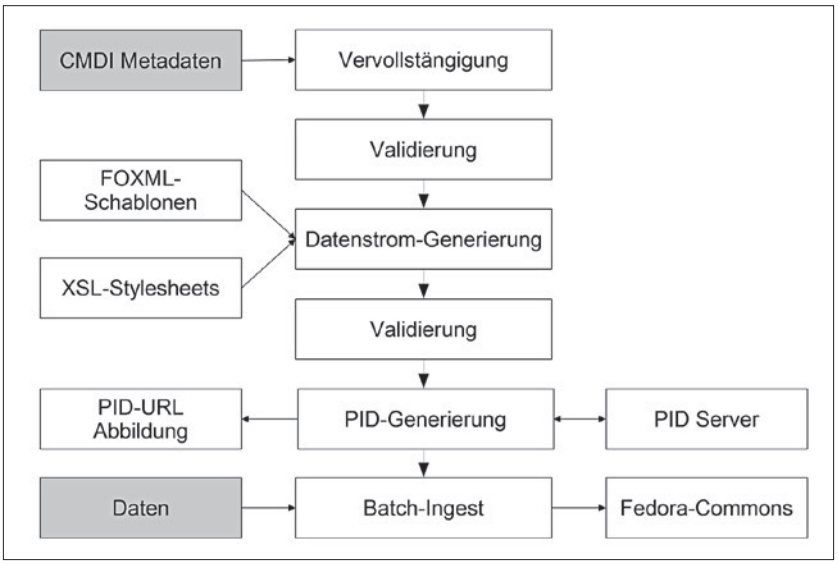

Abb. 5: Generierung von FOXML

Ausblick: Herausforderungen und Perspektiven

Rechtliche Aspekte von Forschungsdaten

Neben rein technischen Aspekten des Forschungsdatenmanagements werden in den beschriebenen Forschungsprojekten auch rechtliche Themen im Umgang mit diesen Daten behandelt. Für Forschungsdaten greifen insbesondere die Regelungen des Datenschutzes und das Urheberrecht. Während datenschutzrechtliche Fragestellungen insbesondere dem Schutz personenbezogener Daten dienen, greift das Urheberrecht sowohl im Falle der Erzeugung als auch bei der Nutzung von Daten (Ludwig & Enke 2013).

personenbezogene Daten

Werden bei der Erstellung von Forschungsprimärdaten *personenbezogene Daten* wie Name, Geburtsdatum, Adresse, Herkunft u. a. erhoben – im Falle des IDS lassen sich insbesondere bei den Aufzeichnungen des gesprochenen Deutsch Rückschlüsse auf den Sprecher ziehen – greifen zunächst das allgemein verbindliche Bundesdatenschutzgesetz (BDSG) und die ergänzende Rechtsprechung der Länder. All diese Gesetze gestatten die Erhebung personenbezogener Daten nur im Rahmen gesetzlichen Vollzugs oder aber unter der Voraussetzung der Einwilligung der betroffenen Person.

Da es Administratoren in der Regel nicht gestattet ist, solche Informationen ohne ausdrückliche Zustimmung der Betroffenen zu speichern, zu verwalten oder gar weiterzugeben, erarbeiten wissenschaftliche Projekte zumeist Richtlinien, die eine Nutzung solch sensibler Daten ausschließlich im Rahmen des angegebenen Zwecks der Forschung vorsehen. Diese Daten müssen während ihres Life Cycle so früh wie möglich anonymisiert werden. Geschieht dies nicht zu Beginn jedweder Nutzung, sind die Möglichkeiten der Verknüpfung der Daten mit anderem Material auf die engen Grenzen des vorgegebenen Forschungszwecks eingeschränkt. Dies bedeutet dann konkret meist, dass eine Publikation dieser Daten ausgeschlossen ist (Ludwig & Enke 2013, S. 56).

Lizenzmodelle

Sofern auch Forschungsdaten den Maßgaben des persönlichen Schaffens, der wahrnehmbaren Formgestaltung, des geistigen Gehalts und der eigenpersönlichen Prägung genügen und die bekannte Frist von 70 Jahren nach dem Tode des Urhebers noch nicht verstrichen ist, unterliegen sie nach deutschem Recht grundsätzlich dem Urheberrecht – unstrukturierte Rohdaten (etwa nicht annotierte und in eine Datenbank überführte Messdaten, vgl. Lutz 2009) sind von dieser Regelung zumeist ausgenommen (Spindler & Hillegeist 2009, S. 23 ff. u. 27 ff.). Als Beispiel sei der Fall der Vervielfältigung von Daten in Datenbanken angesprochen, findet doch diese Form des Datenmanagements in den Geisteswissenschaften inzwischen hohe Verbreitung. Für die Vervielfältigung von in einer Datenbank strukturierten Daten – nicht des softwarebasierten Managementsystems – greift das Urheberrecht nur, wenn es sich um die Verbreitung eines zu definierenden wesentlichen Teils (Spindler & Hillegeist 2009, S. 60) der Daten handelt (dies gilt auch bei einer Verbreitung zu privaten oder Lehrzwecken) (Ludwig & Enke 2013, S. 57; de Cock Bunning et al. 2011, S. 14, 23 ff.; Spindler & Hillegeist 2009, S. 33 ff. und 53 ff.).

Urheberrecht

Unterliegen erhobene Daten also den Voraussetzungen, die das Urheberrecht greifen lassen, sollten wissenschaftliche Institutionen und Projekte dies etwa bei Kooperationen mit anderen Partnern, bei der Publikation oder bei Arbeitsverträgen mit den hauptverantwortlichen Mitarbeitern (Urhebern) berücksichtigen. Ist dies nicht geregelt oder liegt das Einverständnis des Urhebers nicht vor, sind nur bestimmte (Nach-)Nutzungsszenarien denkbar. So ist lediglich eine Vervielfältigung inhaltlicher Fakten im Rahmen eigener Darstellung und Auslegung gestattet. Eine werkgetreue Vervielfältigung ist hingegen lediglich zu privatem und nicht-kommerziellem Gebrauch zulässig. Darüber hinaus gelten die Maßgaben wissenschaftlichen Zitierens. Eine Archivierung von Fremddaten unklarer Rechtslage sollte daher stets mit dem Rechtsinhaber abgestimmt werden (s. de Cock Buning 2011 et al., S. 15 ff. sowie Spindler & Hillegeist 2009, S. 46 ff.).

Weitergabe und Publikation zum freien wissenschaftlichen Gebrauch

Um eine rechtlich abgesicherte Archivierung bis hin zur Verfügbarmachung zu gewährleisten, ist ein Rückgriff auf gängige Lizenzmodelle empfehlenswert. Insbesondere offene Lizenzen, die eine Nutzung unter dem Copyleft-Prinzip vorsehen – die Weitergabe unter ausschließlich denselben Lizenzbedingungen – erscheint sehr von Vorteil (Ludwig & Enke 2013, S. 58 f.; de Cock Bunning et al. 2011, S. 17 ff.). Grundsätzlich ist im Umgang mit solchen Rechtsfällen neben eigener Recherche – das oben beschriebene Projekt CLARIN-D bietet einen juristischen Help Desk für solche Fragestellungen an – die Konsultation eines einschlägig bewanderten Rechtsbeistandes der letztendlich sicherste Weg, um Streitigkeiten mit Urhebern zu vermeiden.

Angesichts der geschilderten Problematik entstehen dem IDS ebenfalls rechtliche Herausforderungen, die es über lizenzrechtliche Maßnahmen zu lösen sucht. Die im IDS in Gestalt von Korpora zum Teil selbst erzeugten, zum Teil aber auch über Datengeber eingeholten vorliegenden Sprachdaten unterliegen verschiedensten Auflagen. So ist bei manchen Daten nach einer bestimmten Nutzungszeit eine Löschung vorgesehen. Da die Datengeber zumeist kommerziell arbeitende Verlage sind, ist eine eigene gewerbliche Nachnutzung durch das IDS von vornherein ausgeschlossen. Andere Lizenzen untersagen die Weitergabe der Daten auf Speicherplätze außerhalb der Rechtsperson des IDS. Eine redundante Datenhaltung außerhalb einer auf einen einzigen Standort beschränkten Einrichtung wird somit verunmöglicht. Schließlich kann die Ausgabe der Daten sich nur so gestalten, dass dem Nutzer von Arbeitsplätzen des IDS ausschließlich Stichproben zugänglich gemacht werden, die keine Rückschlüsse auf den Gesamtkontext der Daten erlauben. Übersieht man diese Einschränkungen, laufen sie den Grundvoraussetzungen eines nachhaltigen und langfristigen Datenmanagements diametral entgegen.

Die Weitergabe und Veröffentlichung solcher Daten zum freien wissenschaftlichen Gebrauch erzeugt einen fundamentalen Konflikt zwischen dem Recht auf Eigentum und dem Recht auf informelle Selbstbestimmung, der eine Harmonisierung von wissenschaftlicher Freiheit einerseits und der rechtlichen Integrität des Urhebers andererseits sehr erschwert. Das IDS strebt daher an, in seinen Mauern ein Kompetenzzentrum zu etablieren, das sich dieser Problematik zumindest für den Bereich der Sprachdaten annimmt und zufriedenstellende Strategien und Lösungen für Wissenschaftler und Urheber findet.

Wissenstransfer und Forschungsdaten-management

Eng verknüpft mit den Fragen des Urheberrechts sind die Möglichkeiten der Nachnutzung von Methoden und Ergebnissen des Forschungsdatenmanagements jenseits des rein wissenschaftlichen Zwecks. Anträge auf Zuwendung von Drittmitteln sehen seit geraumer Zeit als Kriterium für eine Genehmigung das Potenzial für eine Verwertung des beantragten Vorhabens für weiterführende, nicht nur primär wissenschaftliche Zwecke vor. Aus den Reihen der Förderer wird bisweilen beklagt, dass gerade diesem Aspekt, d. h. der Berücksichtigung nicht ursprünglich intendierter, vielleicht sogar kommerzieller Nachnutzungsoptionen, nicht die nötige Aufmerksamkeit zuteil wird.

Zugang und Nutzung der Daten im IDS unterliegen im Wesentlichen den Übereinkünften der Lizenzverträge und Einwilligungserklärungen zwischen den Urhebern und dem IDS. Alle darüber hinausgehenden, ungeregelten oder allgemeinen Nutzungsrechte unterliegen dem Bundesdatenschutzgesetz sowie dem Urheberrecht. Ähnliches trifft auf die den Datensätzen beigeordneten Annotationen zu, deren Verwendungsmöglichkeiten in gesonderten Lizenzabsprachen mit den jeweiligen Softwareentwicklern spezifiziert wurden. In aller Regel bedeutet dies, dass eine Nachnutzung nur durch die Vertragsnehmer des IDS erfolgen kann.

Zwar stellen die wissenschaftlichen Ergebnisse des IDS im Sinne des Rechts noch keine Produkte dar, dennoch erscheint es aussichtsreich, zumindest zu evaluieren, ob die Fähigkeiten eines geisteswissenschaftlichen Instituts auch für die Produktentwicklung weiterentwickelt werden können. Zu diesem Zweck hat das IDS das Projekt »Verwertung Geist – Verwertung von Forschungsergebnissen in den Geistes- und Sozialwissenschaften« ins Leben gerufen (http://vg.ids-mannheim.de). Das Vorhaben greift für die genannten Bereiche die entsprechenden Vermarktungs-, Verwertungs- und Transferfragen auf, um in einem ersten Schritt Potenziale einer weitergehenden Anwendung von Forschungsergebnissen auszumachen. Der nächste Schritt wäre dann, diese Potenziale mithilfe gängiger Methoden und im rechtlichen Rahmen einer monetären oder nichtgewerblichen Verwertung zuzuführen.

Monetäre Rückflüsse an das IDS wie Lizenzeinnahmen, aber insbesondere auch über nutzungsabhängige Gebühren und Beteiligungen an den Umsätzen der entstandenen Produkte, können Geschäftsmodelle sein. Im Falle des Forschungsdatenmanagements im Rahmen des genannten Projektes wird untersucht, wie die Fähigkeiten und Ressourcen des IDS für nichtakademische, kommerzielle Partner nachgenutzt werden könnten. Vor allem Fachverlage wie die Verleger von Wörterbüchern gelten in der Bevölkerung nach wie vor als die maßgebende Instanz zur Definition der Orthographie des Deutschen, auch wenn sie ihre de facto normierende Funktion seit der jüngsten Reform der deutschen Rechtschreibung verloren haben. Neben den Wörterbüchern publizieren solche Verlage u. a. eine Vielzahl weiterer Print- und digitaler Produkte und Softwarelösungen zur deutschen Sprache. Verlage sind wie die moderne germanistische Sprachwissenschaft für ihre Aktivitäten auf die Nutzung umfangreicher, qualitativ hochwertiger digitaler Sammlungen von Texten aus der Gegenwart angewiesen, welche die Basis sowohl für wirtschaftliche Aktivitäten als auch für die Forschung bilden.

Nachnutzungsoptionen

Der Aufbau und die Pflege qualitativ hochwertiger Korpora ist ein äußerst aufwändiges Unterfangen. Im IDS sind seit vielen Jahren durchschnittlich drei volle Mitarbeiterstellen mit dieser Aufgabe befasst. Prinzipiell könnten die IDS-Korpora auch von anderen, insbesondere auch von wirtschaftlich tätigen Unternehmen, genutzt werden. Hierfür sind die weiteren Marktchancen für diese Leistungen systematisch zu erheben und Markteintrittsstrategien zu entwickeln. Da die wirtschaftlichen Kompetenzen nicht in das Kernaufgabenfeld des IDS gehören, gleichzeitig aber – neben den genannten – weitere Verwertungspotenziale gesehen werden, müssen solche Geschäftsmodelle durch im spezifischen Marktumfeld erfahrene Experten beurteilt oder mitgestaltet werden. Die Verlage haben in diesem Kompetenzfeld nahezu ein Alleinstellungsmerkmal – sowohl für den nationalen Bereich als auch für die Beurteilung internationaler Geschäftsmodelle. Nur hierdurch können Probleme bearbeitet werden, die der Verwertung bisher im Wege stehen. Zu diesen gehören lizenzrechtliche Fragen, da die IDS-Korpora meist nur für die Forschung und nicht für wirtschaftliche Zwecke genutzt werden dürfen. Die Hinzuziehung der Expertise solcher kommerzieller Partner böte die seltene Chance, gemeinsam mit einem Verlag Modelle durchzuspielen, wie derartige Verwertungen praktisch umgesetzt werden können.

Projekt »Verwertung Geist«

Dieses Beispiel bietet durch seine Generalisierbarkeit eine einzigartige Chance, auch wenn auf den ersten Blick die Domäne Sprachdaten und die Beteiligten Duden und IDS die Untersuchung eines sehr speziellen Einzelfalls erwarten lassen könnten. Wenn jedoch von der Domäne und dem Einzelfall abstrahiert wird, zeigt sich der prototypische Charakter: Nahezu alle Geistes- und Sozialwissenschaften benötigen qualitativ hochwertige und verlässliche empirische Daten als Basis ihrer Forschungstätigkeit. Ihre Erhebung,

Generalisierbarkeit

Aufbereitung und Bereitstellung ist mit hohen Kosten verbunden. Die kommerzielle Verwertung dieser Daten würde sowohl den Forschungseinrichtungen nutzen, da Geld für die Erweiterung und Verbesserung der Daten zurück fließen würde, als auch den wirtschaftlich tätigen Unternehmen, die Produkte wie z. B. Bücher oder Softwareanwendungen auf Basis einer sehr hochwertigen Datengrundlage erschaffen werden. Eine derartige Datengrundlage kann von einzelnen Firmen meist nicht in ökonomisch vertretbarer Weise entwickelt werden. Diese Form der Leistungsverwertung ist also nicht nur für nationale Medienhäuser bzw. Wörterbuchverlage, sondern eben auch für internationale Unternehmen sowie Anbieter von Online-Informationssystemen interessant. Insofern eröffnet sich ein breites Marktspektrum. Die geisteswissenschaftliche Kompetenz trifft auf einen unternehmerischen Bedarf.

nachhaltige Workflows

Zusammenfassung: Fachtypische Spezifika des Forschungsdatenmanagements

Der tiefgreifende Wandel der letzten Jahrzehnte in den Kommunikations- und Informationstechnologien blieb nicht ohne Folgen für das wissenschaftliche Arbeiten. Auch in den konservativen Geisteswissenschaften werden Forschungsergebnisse zusehends in Gestalt digitaler Primärdaten produziert. Folgeforschungen fußen immer stärker auf Empirie und gründen sich zu erheblichen Teilen auf diesem Material. Eine Aggregation von Forschungsdaten unterschiedlicher Provenienz in ihrer gemeinsamen Nutzung bergen das Potenzial, losgelöst von ihrem ursprünglich intendierten Entstehungszweck weiterführende Forschung zu beflügeln. Somit entzöge der Verlust oder auch nur die Nicht-Interpretierbarkeit von Daten der Wissenschaft immer mehr den Boden für ihre Tätigkeit. Die Implementierung und Aufrechterhaltung einer veritablen Infrastruktur des Forschungsdatenmanagements ist somit nicht nur für die Geisteswissenschaften der nächste konsequente Schritt zu wissenschaftlicher Exzellenz. So vielversprechend die Chancen auf diesem Weg auch sein mögen, so hoch sind die Herausforderungen, den Ansprüchen der wissenschaftlichen Nutzer an die Stabilität ihres neuen Arbeitsplatzes gerecht zu werden.

Handlungsbedarf

Es gehört inzwischen zum guten Ton, Forschungsdaten aus Fördervorhaben der öffentlichen Hand im Sinne des Open Access frei, langfristig und qualitativ abgesichert zur Verfügung zu stellen, um sie einer weiteren wissenschaftlichen Nutzung zuzuführen und jederzeit hinsichtlich ihrer wissenschaftlichen Verlässlichkeit und geordneten Erzeugung verifizierbar zu halten. Dies kollidiert bisweilen mit der guten wissenschaftlichen Praxis, den Schutz der persönlichen, rechtlichen und wissenschaftlichen Interessen von Nutzern, Urhebern und ggf. Probanden anzumahnen. Einem daraus resultierenden eingeschränkten Nutzungsverhältnis müssen Forschungsinfrastrukturen mithin auch genügen. Forschungsinfrastrukturen sollten hierfür einen geeigneten technischen Ansatz bieten, um für essentielle Forschungsdaten jederzeit Erreichbarkeit, Verfügbarkeit, Auffindbarkeit, Reproduzierbarkeit und Referenzierbarkeit sicherzustellen. Die Kosten der Wiederherstellung oder gar Neuerhebung der Daten, die bei Ermangelung entsprechend nachhaltiger Workflows in beständiger Weise auf Forschungsinstitutionen zuzukommen drohen, sind auf die Dauer nicht tragbar.

Open Access

Die unter dem Terminus »Geisteswissenschaften« versammelten Fachdisziplinen sind hinsichtlich ihres Zuganges zu dieser Problematik ausgesprochene Individuen. Auch das hier im Ansatz skizzierte Beispiel der linguistischen Germanistik ist zwar modellhaft, aber nicht in jedem Fall übertragbar. Insbesondere die gemeinsame oder interdisziplinäre Nutzung von Forschungsinfrastrukturen stellt hohe Anforderungen an deren Flexibilität bei der Bewältigung unterschiedlicher Werkzeuge und Datenformate. Soll sich dieser Ansatz künftig in den Geisteswissenschaften bewähren, sollten in Methoden und Quellen vergleichbare oder kompatible Communities insbesondere bei Lösungen zur Langzeitarchivierung von Forschungsdaten die Verwendung von Standards im Format und der Anreicherung von Metadaten in der Dokumentation stets im Blick behalten.

Um diesen Herausforderungen adäquate Antworten zu geben, besteht auf zahlreichen Feldern Handlungsbedarf, der hier nur im Ansatz angesprochen werden kann:

- Forschungsinfrastrukturen benötigen entweder innerhalb einzelner Institutionen oder als vernetzter Ansatz einen geeigneten organisatorischen Rahmen, der ihren nachhaltigen Bestand sichert.
- Eine nachhaltige Finanzierung (Geschäftsmodelle) der Forschungsinfrastrukturen ist unabdingbar. Die Entwicklung neuer, dem technischen Stand entsprechender Tools und Services kann meist nur in befristeten Projekten erfolgen und ist langfristig kaum aus festen Etats zu bestreiten. Die Folgekosten des technischen und organisatorischen Betriebs solcher Infrastrukturen sind in der Forschungsförderung bisher noch nicht ausreichend berücksichtigt. Zudem wirft die Hochschulgesetzgebung (z. B. Art 143c GG) Hindernisse bei der länderübergreifenden Nutzung von Forschungsinfrastrukturen auf. Stets die föderale Verfasstheit

Deutschlands beachtend ist hierbei ggf. eine Anpassung der nationalen Förderpolitik zu erwägen. Eine Option bestünde in der Umleitung von Fördermittelanteilen von Antragstellern hin zu einer institutionellen Grundförderung gemeinsam genutzter Infrastrukturen. Programmpauschalen/Overhead zur Deckung der Folgekosten in Nutzung, Betrieb und Support der Infrastrukturen böten sich demnach an (Riley & Becker 2009b, S. 13 ff.).

- Lösungen auf technischer Basis zur nachhaltigen Verfügbarkeit der Daten sowie die Einbindung neuer Technologien, Methoden und Anforderungen stehen auf der Agenda der nächsten Jahre (Referenzarchitekturmodell, Austauschformate, Standards, Metadaten).
- Es ist zu prüfen, inwieweit für grundlegende Hardware in einer Community-gemeinsamen Nutzung Synergien gebildet werden können. Möglichkeiten, ein geeignetes Netzwerk sowie Identitätsmanagement zu nutzen, sind bereits jetzt gegeben (Eduroam, DFN-AAI).
- Den noch ausstehenden Lösungen zu Fragen des Urheberrechts ist verstärkte Aufmerksamkeit zu widmen. Authentifizierungsinfrastrukturen sollten die angestrebten standardisierten Nutzungsbedingungen und Lizenzen für den Umgang mit Forschungsdaten konsequent berücksichtigen.

Literatur

Broeder et al. 2011 Broeder, Daan; Schonefeld, Oliver; Trippel, Thorsten; Van Uytvanck, Dieter; Witt Andreas: »A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI)«. In: Proceedings of Balisage : The Markup Conference 2011. Balisage Series of Markup Technologies, Vol. 7.

de Cock Bunning et al. 2011 de Cock Buning, Madeleine; van Dinther, Barbara; Jeppersen de Boer, Christina G.; Ringnalda, Allard: Report on the Legal Status of Research Data in the Knowledge Exchange partner countries : Annex 3 : The legal status of research data in Germany. Centre for Intellectual Property Law (CIER), The Netherlands, www.knowledge-exchange.info/Default.aspx?ID=461, 2011.

Fankhauser et al. 2013 Fankhauser, Peter; Pfefferkorn, Oliver; Witt, Andreas (2013): »From TUSTEP to TEI in Baby Steps«. TEI Conference and Members Meeting 2013: October 2-5, Rome.

Francopoulo 2006 Francopoulo, Gil; Declerck, Thierry; Monachini, Monica; Romary, Laurent (2006): The relevance of standards for research infrastructures. Proceedings of International Conference on Language Resources and Evaluation – LREC 2006, Gênes/Italie, 2006. http://hal.inria.fr/inria-00121474.

Ide et al. 2000 Ide, Nancy; Bonhomme, Patrice; Romary, Laurent: XCES: An XML-based Standard for Linguistic Corpora. Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, 2000, pp 825–830.

Ide & Romary 2007 Ide, Nancy; Romary, Laurent: Towards International Standards for Language Resources. In: Evaluation of Text and Speech Systems, Dybkjær, Laila; Hemsen, Holmer; Minker, Wolfgang (Hrsg.), pp 263–284, Springer Netherlands, 2007.

Keitel & Schoger 2013 Keitel, Christian; Schoger, Astrid: Vertrauenswürdige digitale Langzeitarchivierung nach DIN 31644. Berlin: Beuth, 2013.

Lagoze et al. 2005 Lagoze, Carl; Payette, Sandy;Shin, Edwin; Wilper, Chris: Fedora : An Architecture for Complex Objects and their Relationships. In: Journal of Digital Libraries Special Issue on Complex Objects, Springer, abs/cs/0501012 (2005).

Ludwig & Enke 2013 Ludwig, Jens; Enke, Harry (Hrsg.): Leitfaden zum Forschungsdaten-Management: Handreichung aus dem WissGrid-Projekt. Glückstadt: Hüsbusch, 2013.

Lüngen & Sperberg-McQueen 2012 Lüngen, Harald; Sperberg-McQueen, Michael: A TEI P5 Document Grammar for the IDS Text Model. In: Journal of the Text Encoding Initiative (2012), H. 3. http://jtei.revues.org/508

Lutz 2009 Lutz, Peter: Grundriss des Urheberrechts. C. F. Müller, Heidelberg 2009, Rn. 37–86d.

nestor 2010 Neuroth, Heike; Oßwald, Achim; Scheffel, Regine; Strathmann, Stefan; Huth Karsten: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. www.nestor.sub.uni-goettingen.de/handbuch/index.php

nestor 2012 Neuroth, Heike; Strathmann, Stefan; Oßwald, Achim; Scheffel, Regine; Klump, Jens; Ludwig, Jens: Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme. www.nestor.sub.uni-goettingen.de/bestandsaufnahme/index.php

OAI-PMH »The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). www.openarchives.org/OAI/openarchivesprotocol.html

OAIS 2012 Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012.

Riley & Becker 2009a Riley, Jenn; Becker, Devin: Glossary of Metadata Standards. 2009 www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf

Riley & Becker 2009b Riley, Jenn; Becker, Devin: Seeing Standards : A Visualization of the Metadata Universe. 2009. www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards.pdf

Romary 2011 Romary, Laurent: Stabilizing knowledge through standards – A perspective for the humanities. In: Going Digital. Evolutionary and Revolutionary Aspects of Digitization. Herausgegeben von Karl Grandin. Science History Publications, 2011. http://hal.inria.fr/inria-00531019

Spindler & Hillegeist 2009 Spindler, Gerald; Hillegeist, Tobias (2009): KoLaWiss Project : Arbeitspaket 4 – Recht. Göttingen. http://kolawiss.uni-goettingen.de/projektergebnisse/AP4_Report.pdf

Stührenberg 2007 Stührenberg, Maik: Texttechnological Standards – An Overview. In: Andreas Witt, Georg Rehm, und Lothar Lemnitzer (Eds.): Tagungsband : Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007. pp 157–166. Tübingen: Gunter Narr Verlag.

Stührenberg 2012 Stührenberg, Maik: The TEI and Current Standards for Structuring Linguistic Data. In: Journal of the Text Encoding Initiative, Issue 3, October 2012. doi:10.4000/jtei.523. http://jtei.revues.org/523

Stührenberg et al. 2012 Stührenberg, Maik; Werthmann, Antonina; Witt, Andreas: Guidance through the standards jungle for linguistic resources. Proceedings of the LREC-12 Workshop on Collaborative Resource Development and Delivery, Istanbul, Turkey, May 2012.

Sun 2001. Sun, Sam X.: Establishing Persistent Identity Using the Handle System. 10th International World Wide Web Conference, Hong Kong, May 2001.

WissRat 2011 Wissenschaftsrat: Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften (Drs. 10465-11), Januar 2011. www.wissenschaftsrat.de/download/archiv/10465-11.pdf

1 http://hdl.handle.net/10932/00-01B8-AE41-41A4-DC01-5
2 www.isocat.org
3 repos.ids-mannheim.de

Die Verfasser

Dr. Peter Fankhauser, Wissenschaftlicher Mitarbeiter im Programmbereich Forschungsinfrastrukturen, Institut für Deutsche Sprache (IDS), R 5, 6–13, 68161 Mannheim,
E-Mail: fankhauser@ids-mannheim.de

Dr. Norman Fiedler, Wissenschaftlicher Mitarbeiter im Programmbereich Forschungsinfrastrukturen, Institut für Deutsche Sprache (IDS), R 5, 6–13, 68161 Mannheim,
E-Mail: fiedler@ids-mannheim.de

Dr. Andreas Witt, Leiter des Programmbereichs Forschungsinfrastrukturen, Institut für Deutsche Sprache (IDS), R 5, 6–13, 68161 Mannheim,
E-Mail: witt@ids-mannheim.de